The Network Parameter Outlier (NPO) Algorithm

Mark R. Bower

2020-11-10

Problem

Most attempts to relate electrical signaling in a nervous system to the sensations and responses generated by that nervous system (i.e., Systems Electrophysiology) center on the detection, identification and clustering of transient electrical events (e.g., action potentials from individual neurons, ripples during sleep, inter-ictal spikes in epilepsy). These events span multiple spatial (µm to mm) and temporal (µsec to msec) scales and some can only be measured with specific types of electrodes (e.g., action potentials must be recorded with electrodes whose diameter is normally less than 40 µm). Each type of event has a characteristic range of amplitude, duration, frequency and shape (Figure 1).

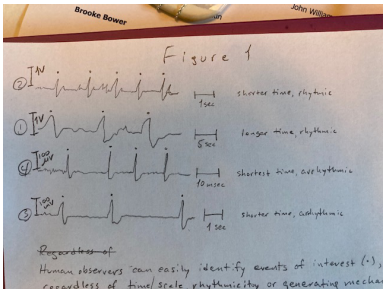

Figure 1: Example electrophysiological waveforms: (from top) heartbeat, respiration, action potential (AP) and interictal spike (IIS). Note that, though the time and voltage ranges differ for each, there are similarities between all of them that are easily detected by the human eye.

Processing electrophysiolgical data generally follows a "pipeline" approach: spectral filtering, threshold detection, feature extraction and clustering. Spectral filtering reduces background noise and removes DC offsets. Threshold detection identifies transient "peaks" in the data that allow time-bounded windows of data to be extracted. Historically, threshold detection was a vital step in long-term recordings, because it reduced the amount of required disk space by ignoring long periods of time where no transient events occurred. It was considered a waste of (expensive and limited!) disk space to just record "noise", but decreasing costs and increasing size of memory now allow all data to be recorded, continuously. Feature extraction encodes each event as a vector of measurements of the detailed features of event waveforms (e.g., peak amplitude, valley amplitude, duration, principal components). These "feature vectors" are then passed to a wide of array of proven clustering algorithms (e.g., Maximum Likelihood Estimators, Support Vector Machines, Neural Networks) in the hopes of identifying and grouping similar events, which are then assumed to have been generated by the same mechanism, such as grouping action potentials thought to be generated by the same neuron. Virtually all Systems Electrophysiology experiments perform these common steps and hypothesis-specific analysis begin at the completion of this pipeline. Any advances in these early signal processing algorithms would, therefore, find broad application across the entire field. While this pipeline approach has been used successfully for decades to advance our knowledge of neural function (the research honored by the 2015 Nobel Prize for Medicine and Physiology utilized these techniques), there are sev-

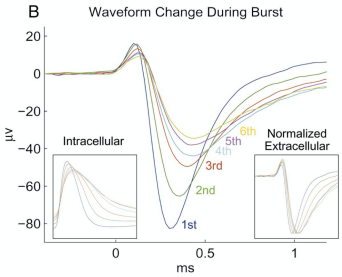

Figure 2: Changing waveform shape from the same neuron during a 'burst' of action potentials as observed by electrodes both inside ('intracellular') and outside ('extracellular') the neuron. These recordings were obtained from an extracted slice of neural tissue, which is the only practical way to record neural activity intracellularly. The occurrence of action potentials are clear in intracellular recordings, providing an 'answer key' for the simultaneous extracellular recording (which are much easier to obtain, but far noisier) and subsequent clustering. This shows how the waveform recorded extracellularly from the same neuron can change dramatically even within a few milliseconds. (Harris et al., 2000)

eral problems that have limited analyses and slowed advances. For one, while many fully-automated detection and clustering algorithms have been created, no fully automated algorithms have gained wide acceptance. Human intervention and oversight are still required to produce the best results. This "semi-autonomous" approach remains the gold standard for all systems electrophysiology, which introduces a time- and labor-intensive step into all data analysis. Another problem arises from the non-stationary aspect of biological signals: detailed features of biological waveforms routinely change with each event and over time, both abruptly (within seconds) and slowly (over days and months) for any of severael reasons (e.g., electrodes moving relative to surrounding tissue, because the electrodes become coated with biological material). Waveform shape can also change in unpredictable ways due to environmental context (e.g, temperature, neuromodulating chemicals, a "burst" of activity, see Figure 2). Standard clustering techniques normally assume that signals are statistically stationary; clustering techniques that do not are less developed and tested. Perhaps the most limiting aspect of this approach, however, is that the computational time required for common clustering algorithms scales faster than linearly with the amount of data. Currently, the computational time required for the fastest clustering algorithms increases at least as $O(n \cdot log(n))$, where n is the number of waveforms to be clustered. This has limited most experiments in this field to a few hours or days of recorded data, at most. Ideally, an algorithm would be found whose computation time scales as $O(n)$; i.e., linearly with the amount of data.

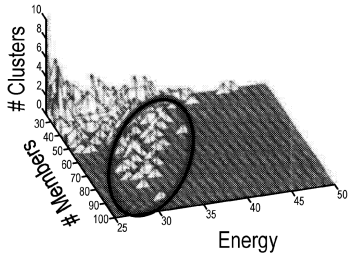

Figure 3: Plot of clusters of action potentials by energy, number of members in each cluster and number of clusters that have those two values. Signals separate from noise because they are rare, repeatable and reliable.

Approach

A completely different approach to "clustering" is used by humans when learning to "read" electrophysiological recordings. Students watch (and sometimes listen to!) recordings with an expert who notes both examples of "noise" (which the student learns to ignore) and "signal" (which the student learns to recognize). Regardless of the type of event, biological signals normally share three properties that differentiate them from noise: they are **rare** (compared to noise), they are **repeatable** (they have roughly the same waveform) and they are **reliable** (their frequency of occurrence is bounded - they occur neither too slowly nor too quickly). Of note, none of these "three R" properties is dependent on the detailed shape of the waveforms. In fact, we become suspicious that a group of signals might actually be artifact masquerading as signal when their signals show little or no variability in shape. Plotting properties that emphasize rare, repeatable and reliable properties of all detected peaks displays a subset of signals

that separate from a larger group; i.e., separates signals from noise. We have coined the phrase *"The Noise Outlier approach"* to describe the use of the noise distribution to help identify the outlying signal distribution (Figure 3).

Algorithm

Because the Noise Outlier approach analyzes data sequentially, similar to how humans address this task, the approach lends itself to graphical representation. When *any* local peak in the data stream is detected, that event can be represented as a node in a graph (Figure 4). Nodes are connected by edges. Each edge takes the value of the correlation coefficient (CC) between the waveforms of the two nodes as its "weight". As events are added to this graph, a difference in the connectivity of nodes within the graph emerges: nodes that represent similarly-shaped rare, repeatable and reliable waveforms (i.e., "signals") form preferentially connected sub-networks that differ from nodes representing common, random waveforms (i.e., "noise"). One way to visualize this takes advantage of the "repeatable" aspect of signals: among the set of "signal" waveforms, one waveform will be closest to the most characteristic shape and will have high CC with a large number of other *signal* nodes. That will cause it to be connected more like a network "hub". Nodes representing *noise*, however, have no such repeatable waveform; the subset of noise signals is generated by random processes that result in a more uniformly connected network. Existing, proven community detection algorithms for graphs (e.g., Girvan-Newman, Louvain) can identify both types of sub-networks in fixed time and existing, proven measures of network connectivity properties can separate *signal* from *noise* sub-networks. We have coined the phrase *"The Network Properties Outlier"* (NPO) algorithm (Figure 5). Because this algorithm runs on a moving window of data with a fixed size it has a time complexity of $O(n)$. In addition, because analysis proceeds sequentially through data, it is ideally suited for real-time data processing.

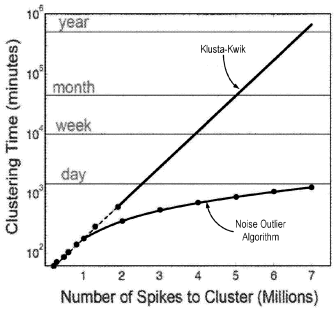

Figure 4: The Network Parameter Outlier (NPO) algorithm clusters data in linear time compared to a current, widely used algorithm (Klusta-Kwik, 'KK') that requires exponential time. Dots represent measured time on real data. The solid line for KK represents an extrapolation of clustering times from smaller data sets; the clustering time for more than 2 million events was impractical. KK uses a Mixture-of-Gaussians assumption to model clusters, which allows it to estimate clusters that contain multiple lobes for neurons like that shown in Figure 2. The time complexity of KK arises from the use of Maximum Likelihood Estimators to fit Gaussian distributions to the underlying data.

Future Directions

The NPO algorithm clusters in $O(n)$ time, because it works on a moving window of data that is bounded in size and therefore is bounded in computational time, causing graph community identification algorithms to operate in bounded time, too. It is known that clustering performance improves as the duration of this moving window increases, but the asymptote (and thus guidance for setting the duration of the window) remain unknown. Quantification of this duration-merit

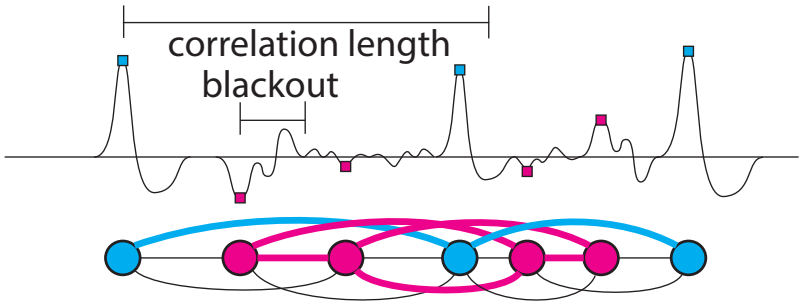

Figure 5: Network Parameter Outlier algorithm. As each peak is detected, a new node is assigned in a graph and connected to existing nodes by edges with weights equal to the correlation coefficient between the associated waveforms. In the schematic, thick (thin) edges represent large (small) weights. Network community detection algorithms find sub-networks based on interconnectivity. Network metrics (e.g., degree, distance, betweeness, centrality, hub score) are computed for each sub-network and used to separate networks whose nodes represent signals (cyan) differentiated from the noise (magenta). The user supplies only two parameters: blackout is the max-min of the expected duration of the signal of interest; correlation length is the expected maximum of time between two events from the same group. Both of these parameters are fixed by physiology for each type of signal. For example, for action potentials (blackout=1 msec, correlation length=100 sec (or 0.01 Hz)); for interictal spikes (blackout=100 msec, correlation length=100 sec). No other parameters are needed.

function requires a computational framework that can handle large graphs, which is provided by the Neocortex computer (Cerberus, Inc.). While direct computation can establish the increase in performance (merit) for increasing data (duration), the optimal solution mechanism (hardware as well as software) remains unclear. Currently, the NPO algorithm has only been run on general-purpose hardware and software, so it is unclear how quickly such an algorithm could run or what the ceiling of its performance might be. Even in the case of using the Neocortex hardware, the physical computational elements were optimized for neural network processing, not graphical network processing. As graphical analysis becomes more popular, WSI of processors and communication optimized for graph processing may become more broadly useful. If so, the NPO algorithm could serve as a starting point for deciding design criteria.